EN GUISE D'AVANT-PROPOS

La rédaction d'un document scientifique pose des problèmes épineux, en particulier dans le domaine de la biologie, en raison de la complexité du langage utilisé pour décrire les processus vitaux.

UN LANGAGE HERMÉTIQUE...

Deux études parues dans la revue *Nature* à 11 ans d'intervalle permettent de mesurer la largeur du fossé qui sépare le langage de la vie courante de celui qu'ont adopté les scientifiques. Un indice appelé **Lex** (pour **Lex**ique) évalue la richesse du vocabulaire employé par différents locuteurs de l'anglais en se référant à un thésaurus de quelque 87 000 mots, recensés dans divers magazines, romans, manuels et encyclopédies. L'indice lexical le plus bas (-59) définit le vocabulaire d'un vacher qui parle à ses bêtes. Une conversation banale entre adultes a une valeur de Lex estimée à -41. Les écrits présentent des difficultés lexicales croissant de -32 pour les livres destinés aux enfants, à -27 pour les bandes dessinées et à -19 pour les ouvrages de fiction réservés aux adultes. Par convention, l'indice 0 est attribué aux journaux quotidiens tels que le *New York Times* et le *Daily Telegraph*.

En matière scientifique, le niveau de complexité lexicale n'a pas cessé de croître pendant plus d'un siècle. Prenons le cas des deux revues les plus importantes consacrées aux sciences naturelles *(Nature* et *Science)*. Leur indice lexical est passé de 0 en 1900 à 15 en 1950, puis à 30-35 en 2000, avec des pointes à près de 60. Cela signifie qu'une personne capable de lire les journaux quotidiens pouvait jusqu'en 1900 comprendre ce qui se publiait dans *Nature* et *Science*. En 1950, c'était difficile. En 2000, c'est devenu impossible.

Les autres publications scientifiques ont évolué de la même manière, mais à des vitesses variables suivant leur spécialité. En matière d'hermétisme, les biologistes battent largement les chimistes, les géologues, les astronomes et les physiciens. Notons cependant que l'indice lexical ne tient pas compte des équations, peu nombreuses dans les écrits relevant de la biologie, ce qui fausse la mesure des difficultés de compréhension.

DES MÉCANISMES COMPLEXES...

La complexité du vocabulaire employé par les biologistes tient à la structure des êtres vivants et à leur manière de fonctionner. Les cellules, et a fortiori les organismes multicellulaires, sont des machines extrêmement complexes, où des milliards de molécules interagissent en permanence, dans un espace confiné. Comme chaque molécule peut interagir plusieurs autres, elle possède souvent plusieurs fonctions. C'est particulièrement vrai pour les protéines, dont les cellules contiennent des milliers d'espèces différentes. Chacune de ces espèces doit recevoir un nom. Chaque fonction doit être décrite par un mot, puisé dans le vocabulaire courant ou créé de toutes pièces.

Les êtres vivants semblent éprouver le besoin de multiplier les interactions qui se déroulent en leur sein et les agents de ces interactions. Cette tendance est due à leur manière d'évoluer, que l'on peut qualifier de conservatrice : ils aiment acquérir, mais détestent jeter. La propension à tout conserver s'explique en partie par le fait qu'aucun organisme ne peut à chaque génération détruire une

trop grande partie de l'héritage que lui ont légué ses prédécesseurs, sous la forme d'un certain nombre de gènes. S'il perd une part importante de l'information contenue dans ceux-ci, il ne peut plus synthétiser les macromolécules (ARN et protéines) qui lui permettraient de fonctionner correctement. Les cellules supportent mal les pertes de mémoire. Mais elles acceptent les gains sans trop de réticences.

Cette tendance au conservatisme se manifeste chez certains animaux qui possèdent une mémoire cellulaire surdimensionnée par rapport à la complexité de leur anatomie et de leur physiologie. Ainsi, le ver nématode *Caenorhabditis elegans*, dont il sera beaucoup question par la suite, a presque autant de gènes que les mammifères. Pourtant, son corps se compose essentiellement d'un système reproductif, d'un système digestif et d'un système nerveux. Il est probable que cet animal dérive d'un ancêtre beaucoup plus complexe qui aurait subi une simplification radicale de son organisation, mais conservé l'essentiel de ses gènes.

DES GÈNES PAR MILLIERS...

La mémoire cellulaire s'enrichit par un mécanisme simple, qui consiste à dupliquer un gène ou un groupe de gènes, à la suite d'une anomalie de la méiose. Le plus souvent, les deux gènes dupliqués se retrouvent côte à côte sur le même chromosome. Si l'opération se répète, elle crée des batteries de gènes contigus.

Si les choses demeuraient en l'état, elles n'entraîneraient pas trop de complications. Mais les gènes dédoublés ne restent jamais identiques. En dépit de la fidélité qui caractérise la réplication du matériel génétique (l'ADN), des changements se produisent de temps à autre. Des mutations surviennent et s'accumulent au fil des générations, si bien que les gènes dupliqués ne cessent d'évoluer. L'un perpétue la fonction du gène primitif, dont l'agent d'exécution est presque toujours une protéine. L'autre peut acquérir une fonction nouvelle, différente de celle qu'assure son compagnon. Par le double jeu des duplications et d'évolution divergente des gènes, les cellules ont peu à peu enrichi leur panoplie de protéines.

La duplication n'est qu'un mécanisme de diversification parmi d'autres. Suivant un mot célèbre de François Jacob, les êtres vivants évoluent en bricolant. Ils savent non seulement dédoubler leurs gènes, mais aussi les remanier de multiples façons : les fragmenter, les retourner, les transloquer, en créer de nouveaux en combinant des pièces de différentes origines, ou encore les allonger en multipliant un motif structural, ce qui donne à leurs produits (ARN et protéines) une organisation répétitive, comportant des séries identiques ou similaires de nucléotides ou d'acides aminés. Les virus participent au remue-ménage général, car ils peuvent s'introduire dans les chromosomes de la cellule qu'ils infectent, puis s'en extraire et emporter un fragment d'ADN, qu'ils implanteront ailleurs.

UNE MÉMOIRE CELLULAIRE SURDIMENSIONNÉE...

Les progrès rapides des techniques ont permis de séquencer en quelques années l'ADN contenu dans le ou les chromosomes de plusieurs dizaines d'espèces bactériennes, végétales et animales. En 2001, deux groupes de chercheurs publiaient une première version de la séquence nucléotidique de l'ADN humain. Pour donner une idée du caractère titanesque de l'entreprise, signalons que pour consigner sur un support matériel l'ensemble des quatre signes (nucléotides) dont l'agencement constitue la mémoire enfermée dans l'ADN de nos 23 chromosomes, il faudrait l'équivalent d'une quinzaine de dictionnaires Robert, incluant au total environ trois milliards de signes. Il n'y a ni coupures, ni espacements dans ces 23 séries interminables de lettres, que les cellules conservent et propagent sous la forme de paires de lignes disposées côte à côte, représentant les deux brins de la double hélice.

Pour repérer dans cette liste la présence d'une information utile, il a fallu mettre en œuvre des algorithmes de recherche puissants, élaborés en déterminant des motifs structuraux typiques de telle ou telle macromolécule. Ces motifs se présentent sous la forme de séquences, dites « consensus » construites en comparant les ARN ou les protéines censés remplir la même fonction chez divers organismes, y compris les bactéries. Ces séquences sont les vestiges d'un passé commun. C'est l'héritage légué à leurs descendants par les organismes fondateurs de chaque groupe, héritage sans cesse modifié au cours des générations successives, mais pas au point de devenir méconnaissable, en dépit d'innombrables erreurs de copie qui se sont accumulées au fil du temps.

D'abord estimé à 30 000, le nombre de gènes humains a été par la suite réévalué à la baisse. Les chiffres les plus récents font état d'environ 21 000 gènes spécifiant des protéines, et un nombre assez mal connu de gènes dont les produits sont des ARN intraduisibles. Le travail de séquençage fait clairement ressortir la tendance générale au conservatisme qui caractérise la transmission de la mémoire cellulaire. Parmi les quelque 21 000 gènes identifiés chez l'homme, beaucoup peuvent être regroupés en familles de taille variable, créées par duplications en cascade. Les familles comportent moins de dix à près de 800 représentants. Les 30 plus vastes regroupent au total presque 8000 membres, qui sont éparpillés sur tous les chromosomes et séparés par d'énormes intervalles, remplis de séquences apparemment dépourvues de signification, dont beaucoup ont été introduites par des virus et consciencieusement recopiées de génération en génération.

Il n'existe pas seulement des espaces « vides » entre les gènes. Il y en a aussi à l'intérieur, puisque la plupart d'entre eux comportent des régions internes, appelées introns, qui ne servent apparemment à rien, parce qu'elles ne figurent pas dans leurs produits finaux (les ARN fonctionnels). Ces derniers ne conservent que les exons, c'est-à-dire les éléments de séquence encadrant les introns. Le nombre d'introns par gène varie de 0 à près de 200, avec une moyenne de 9, et une longueur cumulée 20 fois supérieure à celle des exons. Au total, les séquences vraiment utiles (les exons) représentent environ 1,5 % de l'information véhiculée par l'ADN humain. Mais compte tenu de la longueur des introns, nos cellules peuvent en fin de compte transcrire, c'est-à-dire copier sous la forme d'ARN, environ 30 % de leur information génétique, ce qui représente environ un milliard de nucléotides. Un chiffre énorme.

DES COMPLICATIONS APPAREMMENT SUPERFLUES...

Une impression d'irrationnel se dégage à l'examen de l'organisation des chromosomes et des mécanismes qu'ils gouvernent. Si un ingénieur, plutôt qu'un bricoleur avait dirigé les processus évolutifs, les choses se seraient sans doute passées autrement. Trois exemples, qui seront développés dans le texte qui suit, illustrent la propension, ou plutôt la tolérance à la complication qui caractérise les mécanismes cellulaires.

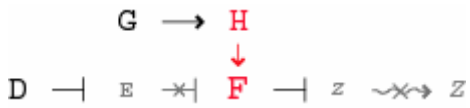
Les mammifères en général et l'homme en particulier ne craignent pas d'encombrer leur génome avec une énorme quantité d'information qui semble dépourvue d'utilité. Mais ils font quelquefois preuve de parcimonie en empilant deux gènes dans un même espace. Dans l'ADN du chromosome 9 humain, une petite région sert à ordonner deux protéines de fonction similaire (INK4a et Arf), mais formées de séquences d'acides aminés complètement différentes. Il semblerait plus rationnel, et en tout cas plus simple, de confier les deux tâches à des gènes distincts, fussent-ils contigus et contrôlés par un mécanisme commun.

Il paraît tout aussi illogique de placer un processus vital - la division des cellules - sous le contrôle de deux gènes antagonistes *(mdm2* et *p53)*, dont les mammifères peuvent se passer sans trop de dommages, puisque l'inactivation de l'un de ces gènes provoque des anomalies létales chez les souris, alors que l'inactivation de l'un et l'autre n'a guère de conséquences. Apparemment, ces animaux ont innové en confiant au couple *mdm2-p53* le contrôle de la prolifération cellulaire, car les invertébrés étudiés jusqu'à présent ne l'utilisent pas, tout au moins à cet usage. Dans une certaine mesure, cette prise de contrôle apparaît comme un luxe inutile.

D'une manière plus générale, les chaînes d'interaction entre les gènes ou leurs produits sont plus compliquées qu'il ne semble à première vue nécessaire. Il arrive souvent qu'un facteur (A) stimule par l'intermédiaire d'une série de relais (B et C) l'activité d'un gène *(y)*, donc la synthèse de la protéine correspondante (Y). Les chaînes d'interaction font fréquemment intervenir des suites d'inhibition et d'activation entre protéines (représentées respectivement dans le texte par un T couché et par une flèche), aboutissant à la mise en activité d'un gène (représentée par une flèche ondulée) :

En fait, un nombre pair d'inhibitions (ci-dessus) correspond à une stimulation, tandis qu'un nombre impair aboutit à une répression :

Ces complications ne sont pas purement gratuites, car elles offrent le moyen de connecter plusieurs chaînes d'interaction :



Mais elles rendent singulièrement ardu le travail des biologistes. Une même cause (par exemple une mutation) ne produit pas toujours des effets constants ni faciles à interpréter, parce que les systèmes étudiés sont hétérogènes et que d'innombrables interconnexions embrouillent leur fonctionnement. Ce n'est pas une raison pour renoncer à la méthode analytique et pour considérer que les processus vitaux, marqués par une indétermination qui rend parfois leur déroulement imprévisible, resteront toujours environnés de mystère.

UN VOCABULAIRE PEU COHÉRENT...

Pour tenter de comprendre et de décrire le fonctionnement de la machinerie cellulaire, les biologistes ont dû développer des méthodes de plus en plus élaborées et créer un vocabulaire de plus en plus complexe, comportant un nombre sans cesse croissant de mots, de sigles et d'acronymes. Des termes nouveaux servent à désigner des techniques récemment mises au point, des molécules et des mécanismes inconnus jusqu'alors. C'est le domaine des molécules qui connaît l'expansion la plus forte. En quelques années, on a découvert des milliers de protéines qu'il a fallu nommer, de même que les ARN et les gènes qui dirigent leur synthèse.

Malheureusement, la terminologie s'est développée de manière peu rationnelle, de sorte que les innombrables néologismes et abréviations qui pullulent dans les publications biologiques n'apprennent souvent rien concernant la structure et les fonctions des entités matérielles ou immatérielles qu'ils désignent. L'existence de synonymes n'est pas non plus de nature à faciliter les choses. Pour couronner le tout, des termes dont la signification paraît évidente peuvent induire en erreur. Ainsi, les facteurs de croissance n'agissent pas toujours chez les animaux en augmentant la taille et le nombre des cellules qui les constituent. Il en est qui ont plutôt tendance à freiner la prolifération cellulaire.

La terminologie génétique est particulièrement confuse. Beaucoup de gènes sont identifiés par le nom d'une mutation qui perturbe leur activité. Dans les cas favorables, ils reçoivent un nom évocateur, facile à retenir. Par exemple, le gène *chico*, qui signifie *petit garçon* en espagnol, réduit la taille des mouches du vinaigre quand une mutation l'empêche de fonctionner correctement. Mais le mot ne précise nullement la fonction du produit du gène (la protéine Chico), qui est de stimuler le métabolisme du glucose, donc la croissance des larves. Tout aussi mystérieuse est la signification des sigles ou acronymes désignant les anomalies qui ont servi à dénommer les gènes *ink4a, arf* et *mdm2*.

En principe, la situation devrait être meilleure quand le nom d'un gène et d'une protéine évoque directement une de leurs fonctions, réelle ou supposée. Mais ce n'est pas nécessairement le cas. Personne ne peut deviner que le produit du gène *ctl* est un enzyme (la catalase) qui décompose le peroxyde d'hydrogène. L'hermétisme franchit un nouveau palier lorsque les protéines n'avaient pas de fonction connue au moment de leur dénomination. Passe encore si l'abréviation ou le sigle adopté fournit un renseignement sur une propriété de la protéine. Par exemple, le gène *p53* spécifie une protéine (p) dont le poids moléculaire est d'environ 53 000. L'opacité devient totale lorsqu'un nom de code ne fournit aucune information qui pourrait éclairer le lecteur. Ainsi, les chiffres désignant le gène *14-3-3* évoquent trois opérations qui ont été nécessaires pour purifier la protéine portant ce nom bizarre.

MULTIPLIENT LES DIFFICULTÉS DE COMPRÉHENSION

Deux options s'offrent aux auteurs désireux de décrire les progrès récents des recherches dans un domaine particulier de la biologie. La première consiste à renoncer au jargon propre à chaque sous-discipline. Mais cela conduit à utiliser des périphrases alambiquées et à remplacer le jargon par un autre, peut-être encore plus lourd. La seconde option est de recourir au jargon, quitte à en expliquer le sens, ce qui oblige souvent à remonter d'article en article jusqu'à la publication originale décrivant la création de chaque sigle ou acronyme. C'est le parti qui a été adopté dans le texte qui suit.

Il est évidemment impossible de définir tous les concepts de base créés par les biologistes (gènes, allèles, dominance, homozygotie, etc.), d'expliquer les termes généraux décrivant la structure et le fonctionnement des cellules (ADN, ARN, protéines, glucides, mitose, méiose, réplication, transcription, traduction, etc.), et de fournir le mode d'emploi des techniques les plus courantes utilisées dans les laboratoires (chromatographie, électrophorèse, clonage moléculaire, séquençage de l'ADN, obtention de protéines pures par génie génétique, etc.). Les explications se limiteront aux termes assez peu usités et aux techniques développées depuis peu. Certains choix pourront sembler arbitraires et certains éclaircissements inutiles ou pas assez explicites. Ces inconvénients sont inévitables, car tous les lecteurs n'ont pas les mêmes centres d'intérêt, ni la même formation scientifique. Tous n'ont pas non plus gardé en mémoire les détails de ce qu'ils ont appris au cours de leurs études ou étudié par la suite.

Le texte qui suit devrait être accessible aux biologistes ayant une formation universitaire. Il propose plusieurs niveaux de lecture ; (**1**) un exposé relativement bref : (**2**) un ensemble d'annexes détaillant des techniques, des mécanismes et des maladies dont la description est nécessaire ou utile à la compréhension générale, mais paraîtra sans doute superflue à certains lecteurs (**3**) une série d'explications, comportant plus de 200 entrées, relatives aux mots ou abréviations marqués par un ou deux astérisques dans le corps principal de l'exposé et dans les annexes.

Après une brève introduction historique, sept chapitres de longueur très inégale traitent des processus qui peuvent limiter la durée de la vie des organismes unicellulaires et pluricellulaires. Les trois premiers ont une vocation essentiellement descriptive. Les trois suivants sont plus spéculatifs. Le dernier traite des moyens qui pourraient retarder la sénescence des cellules, donc éventuellement prolonger la vie humaine. Par la force des choses, le texte ne rend compte que d'une infime partie des connaissances accumulées sur le sujet choisi : entre 50 et 100 articles paraissent *par jour* sur la sénescence et la mort cellulaire. Dans le domaine de la biologie, le nombre total de *publications quotidiennes* dépasse largement le millier. À l'heure actuelle, la liste des références mentionnant la protéine P53 comporte plus de 40 000 entrées...

BIBLIOGRAPHIE

Hayes DP. The growing inaccessibility of science. *Nature* 1992; **356**: 739-40.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860-921.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**: 931-45.

Knight J. Clear as mud. *Nature* 2003; **423**: 376-8.

Venter JC, Adams MD, Myers EW, Li PW *et al*. The sequence of the human genome. *Science* 2001; **291**: 1304-51.